DOCUMENT RESUME

ED 213 754 TM 820 176

AUTHOR Wendler, Cathy L.W.
TITLE Optimizing Test Selection Using a Cost Analysis Approach.
PUB DATE Mar 82
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Bayesian Statistics; *Cost Effectiveness; Methods; *Models; Testing Problems; *Test Selection

ABSTRACT

It is commonly proposed that educational and psychological assessments consist of a multifaceted procedure that incorporates scores from a variety of tests and resources. Little guidance has been offered regarding the method for selecting a combination of testing instruments used in an evaluation. The problem of test selection is discussed in the present paper using a cost-efficiency approach. A Bayesian technique for defining test efficiency is presented as a key component in the cost analysis model. Two methods for interpreting C/E ratios are suggested. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED213754

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Optimizing Test Selection Using a Cost Analysis Approach

Cathy L.W. Wendler

Western Psychological Services
Los Angeles, California

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C.L.W. Wendler
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the meeting of the American Educational Research Association, New York, March 1982. The contribution of Gale H. Roid, Director of Research, Western Psychological Services, is gratefully acknowledged.

Requests for further information should be directed to Cathy L.W. Wendler, Staff Psychologist, Western Psychological Services, 12031 Wilshire Blvd., Los Angeles, California 90025.

TM 820 176

It is commonly proposed that educational and psychological assessments consist of a multifaceted procedure that incorporates scores from a variety of tests and resources. Little guidance has been offered regarding the method for selecting a combination of testing instruments used in an evaluation. The problem of test selection is discussed in the present paper using a cost-efficiency approach. A Bayesian technique for defining test efficiency is presented as a key component in the cost analysis model. Two methods for interpreting C/E ratios are suggested.

Educational decisions are generally based on the program or teaching approach that produces the largest and "best" effect on students, teachers, or administrators. Little concern has traditionally been given to the costs of such programs or approaches, providing that the costs fall within the prescribed limits of funds. Cost analysis, while widely used for policy implementation in fields outside of education, has been largely ignored within educational decisions. This has been the result of a number of factors, including lack of knowledge as to cost analysis procedures by educators and the feeling by many that the most beneficial or effective program should have precedence, regardless of the cost involved. However, with increasing financial stress on schools, the use of educational resources is coming under greater examination. School officials are no longer just concerned with the program that offers the greatest benefits for students and staff but with the program that is best for the funds available. Decisions must be made on how to reduce expenditures or alter educational services to meet new requirements within the constraints of existing educational resources (Levin, 1981). The use of cost analysis in making educational decisions will help define the program that provides the maximum results relative to the available funds, although the program may not be the least expensive nor the most effective (Roid, Note 1).

The advent of several public laws, in particular Public Law 94-142, has made "more-for-your-money" evaluations an important issue. By elaborating as to the age and type of child eligible for special services, schools have developed massive backlogs for

evaluation and find existing special services personnel and resources inadequate (Wright, Padilla, Renneker, & Stearns, Note 2). Many schools have reacted to this dilemma by restructuring evaluation strategies. The result of this may be quicker evaluation and placement decisions, but whether the restructuring has resulted in procedures that produce the most efficient decisions compared to the costs is undetermined. Thus, the question of concern in the proposed cost analysis model is the evaluation procedures that produce the maximum results (i.e., correct classifications of children) relative to costs.

Cost-efficiency vs. Cost-benefit. Cost analysis may be approached using a number of techniques, the most common being those of cost-efficiency (CE) or cost-benefit (CB). Often the terms CE and CB are used interchangeably, although the basic assumptions behind the analyses are quite different. Briefly, CB analysis refers to the assessment of alternatives according to both the costs and benefits that they confer upon society (Levin, 1981). In CB analysis each alternative is evaluated as to the pecuniary value of its resources and its contribution to social benefits; its outcomes are expressed in monetary terms. Alternatives whose benefits exceed their costs are the only ones that should be considered and those alternatives having the highest ratios of benefits to costs should be selected. An example of a CB analysis would be the comparison of several educational programs using different techniques to teach adults automotive skills. Thus, CB analysis is able to generalize across several programs with different goals or objectives.

CE analysis, in contrast, assumes that the alternatives have the same objectives or goals and that a common measure of effectiveness can be used to assess them (Levin, 1981; Roid, in press). Additionally, CE analysis has potentially more value when evaluating alternatives that cannot be converted to pecuniary values (e.g., self-esteem, mathematics test scores). The selected alternatives should be the alternatives that provide the maximum effectiveness per level of cost or that require the least cost per level of effectiveness. Because of its nonmonetary effectiveness measures, CE analysis is thought to have a wider applicability in educational evaluation than CB analysis (Roid, Note 1).

The technique of interest here is CE analysis. The efficiency of evaluation techniques in placement decisions is a key issue in the proposed model; that is, the degree to which classifications of examinees will result in correct placement decisions.

Determining Combinations of Tests. A number of information sources should be used when evaluating a child for placement into a special services program, or when evaluating any examinee for placement into a program. The procedure to be proposed here determines the efficiency of a group of assessment instruments. Since standardized tests have in the past been used for evaluation purposes, and will probably continue to be used, this paper focuses on the use of tests as assessment instruments. However, the model may be expanded to include other information gathering activities, such as teacher observations, interviews with parents, medical histories, and so forth.

The initial step when using the model is to determine which tests might be selected for use in an evaluation. Several factors should be considered in the test selection process:

1. The selection of tests must be based on both ethical and empirical grounds (Messick, 1980). Ethical considerations address the appropriateness of a test by assessing the potential "social consequences" of testing. Empirical considerations deal with psychometric evidence, particularly evidence of construct validity.
2. The tests selected for an evaluation must be logically determined. Instruments should be used that assess actual need areas of an individual. Administering tests that measure sensorimotor integration to a child who actually has a hearing impairment may well be a waste of resources. Thus, the combination of tests must be logically related to the examinee's needs.

Once the dimensions of ethics, empiricism, and logic are considered, combinations of tests suited to a particular evaluation may be created. The question as to which combination of tests is the most cost-efficient may then be addressed.

Definition of Efficiency. As used here, efficiency refers to the probability of evaluation techniques correctly placing individuals into their proper group or classification. The concept of efficiency is used as a comparison of evaluation procedures. For each group of procedures (i.e., tests) there is a value E so that if $E_1 > E_2$, procedures in Group 1 are said to be

better, or more efficient, than procedures in Group 2 (Larsson, Note 3).

The process of determining the efficiency of a group of tests involves a placement problem rather than a mere classification problem. A placement procedure is a special case of classification. In a placement decision a number of dimensions or scores are combined to form a single composite scale; scores on the same composite scale are used to make decisions between placement outcomes (Cronbach & Gleser, 1965). Figure 1 illustrates the nature of a placement problem.

Place Figure 1 About Here

Generally, the efficiency of a test is found by determining the number of subjects classified by the test into each of four possible groups: false positives, true positives, false negatives, and true negatives. The ratio of correctly classified subjects (i.e., sum of true positives and true negatives) to the total number of subjects is said to be the discriminatory efficiency of the test, or:

$$\% \text{ efficiency} = \frac{n_1 + n_2}{N},$$

where n_1 = number of true positives

n_2 = number of true negatives

N = total number of subjects.

However, this simple ratio fails to account for a number of factors. In essence, it does not provide a measure of probability

of correct classification for a test but is a ratio based on a defined sample. The introduction of more or different subjects may dramatically reduce or increase the discriminatory power of the test. The ratio is also limited in scope; that is, when taken alone a test's efficiency may be limited, but when used in conjunction with other tests, it may produce highly accurate placement results. Finally, no prior information about the test or subjects is accounted for in this ratio. For example, it is easier to obtain higher efficiency on a test designed to measure the difference between chronic schizophrenics and neurotics than a test designed to measure the difference between acute schizophrenics and neurotics since there is a greater difference between the thought processes of chronics and neurotics than between acutes and neurotics. Another example is a test that is more efficient when assessing a specific range of abilities; it may be highly accurate when used with average or below average children but inaccurate for gifted children.

One method for confronting the limitations of ratio-determined test efficiency is to modify the computation by introducing Bayes' theorem (Schmitt, 1969). Two basic advantages of Bayesian techniques include (Jones & Newman, 1971): (1) prior information about the problem is used in the decision-making process and (2) a specific probability statement about the results is arrived at through the computations. A number of Bayesian approaches may be considered when addressing the issue of test efficiency. The most appropriate approach here is to determine the efficiency of a test by establishing its probability of correct placement based on the

results of a priori knowledge about the diagnostic category of interest and the results of a validation study or other type of placement study.

Several individuals have used similar approaches for evaluating selection decisions. Jones and Newman (1971) defined a model which established a posterior distribution based on the prior distribution and the sample and then compared the result to an established criterion. They applied their model to the specific problem of test selection. Leroy and Mellergard (1973) approached the problem of diagnostic decisions by using the actual number of subjects in a diagnostic category as the basis for prior probability. Their model simplified Bayes' theorem so that the conditional probabilities were additive. A third Bayesian model, presented by Brumet (Note 4), regarded the posterior probability of an event as a function of initial probability and experimental data. A probability value was an indication of the strength of a belief or opinion about the likelihood of an event occurring.

According to Brumet (Note 4), Bayes' theorem is "essentially an algebraic relationship by which prior probabilities are revised in view of additional data to obtain posterior probabilities" (p. 14). Prior probability is the probability of something occurring before an observation or experiment is conducted. A prior probability, in this paper, is defined as the probability of belonging to a particular diagnostic category. This will be represented by $P(C_1)$. While this probability is subjective, it may also be influenced by some commonly known data. For example, it has been

proposed that possibly 7 percent of the school population require special educational services for learning disabilities (Kirk, 1972). Thus, the probability that a child would belong to a learning disabled group would be $P(C_1) = .07$. This is called a "base rate" for a population characteristic. The P(C) statement, therefore, may be based on a "frequency" concept of probability, past experience, or simply a personal feeling (a "hunch").

The probability of correct classification following data collection or observation will be represented by $P(D/C_1)$. This is the probability of a diagnostic indicator (e.g., pattern of test scores) being positive*, given that the subject actually belongs to a particular classification (or D given C). Generally, this probability estimate will be the result of a validity or placement study that determines the percent efficiency of a group of tests. Each test within the group of tests provides some judgment (e.g., diagnosis, test score, ranking) as to which diagnostic category an examinee belongs. For example, in a group containing three different tests, a number of possible outcomes are available:

(1) The individual may receive subaverage scores on all three tests, designated as PPP, meaning the examinee was diagnosed as positive by all three tests.

(2) The individual may receive average (or above average) scores on all three tests, designated as NNN, meaning

---

*This may be more easily understood using the medical approach to positive and negative; that is, positive on a test indicates the presence of a disability while negative indicates its absence.

the examinee was diagnosed as negative by all three tests.

(3) The individual may receive a combination of scores: PPN, PNP, or NNP.

$P(D/C_1)$ is the number of individuals belonging to the diagnostic category who are classified into that category by the group of tests (e.g., PPP). This is best conceptualized as a frequency table which indicates the actual group classification of an individual and the diagnosed classification based on the results of the tests.

Place Figure 2 About Here

Since the category of interest here is the probability of belonging to the "subaverage" category, the cell to be used in determining $P(D/C_1)$ is that cell containing the true positives (Cell #1); that is, the number of individuals actually belonging to Group A who are diagnosed as belonging to that group by all tests. $P(D/C_1)$ may be found using the frequency found in Cell #1 by:

$$P(D/C_1) = \frac{n_1}{N_1},$$

where n_1 = number of individuals in Cell #1

N_1 = number of individuals in Group A.

Posterior probability, $P(C_1/D)$, provides the efficiency statement for the group of tests. This probability is based both on the prior probability $P(C_1)$ and the results of the

data collection process $P(D/C_1)$, so that:

$$P(C_1/D) = \frac{P(D/C_1)\ P(C_1)}{P(D)},$$

where $P(D) = P(D/C_1)\ P(C_1) + P(D/C_2)\ P(C_2)$.

$P(D/C_2)$ is defined as the probability of being placed into Classification A when the examinee actually belongs to Group B (i.e., being placed into Cell #2). It is concerned with the number of individuals in Group B who are diagnosed by <u>all</u> tests as belonging to Classification A. $P(D/C_2)$ may be found, using the frequency of Cell #2, by:

$$P(D/C_2) = \frac{n_2}{N_2},$$

where n_2 = number of individuals in Cell #2

N_2 = number of individuals in Group B.

$P(C_2)$ is the probability of <u>not</u> belonging to Group A but of belonging to Group B, or:

$$P(C_2) = 1 - P(C_1).$$

<u>Definition of Costs.</u> The second key component in the proposed model considers the costs of tests used in evaluations. Schools could reasonably try to obtain a wider variety of information on a student by increasing the number of tests used during an evaluation while attempting to stay within budgetary constraints. Costs refer to options or opportunities lost (Fisher, 1971; Levin,

1981). For example, purchasing one particular combination of tests would prevent a school from purchasing another combination of tests; thus, costs would be attributed to the first combination of instruments. However, defining what comprises true costs in education is difficult because school factors are not always readily separated from factors not directly related to the production of knowledge (Bible, 1976). One way to arrive at a cost estimation is to restrict the definition of costs to measure only those things that are easily measured (Fisher, 1971).

Any number of costs may be allocated to a single test: administration costs, material costs, building and room costs, maintenance costs, and so on. But not all of these should be considered as feasible costs involved with giving a single test; it is importance to distinguish between relevant and irrelevant costs. According to Fisher (1971), relevant costs lie in the future, not in the past. Past costs, or sunk costs, might include such things as building and room costs. However, since the building and room probably already existed and were not built just for the purpose of testing, they are not considered relevant costs. Additionally, it is important to obtain costs involved for the particular time period in which costs are being considered, or annualized costs (Levin, Note 5). Further, a test with all consumable materials would cost more than a test with reusable materials, since materials would need to be replaced following each evaluation. It is also possible that several tests might share the same materials (e.g., writing materials) and reduce the cost per test for the materials. These costs are said to be

joint costs (Levin, Note 5). For the purposes of this paper, however, costs are defined in a more simplistic manner. It is recognized, nonetheless, that many variables need to be considered in order to decide what comprises true costs of an evaluation instrument.

Costs here are defined as the cost of the testing materials and the costs associated with the test's administration. Material costs are on a per student basis. If the cost of a set of testing materials is $25.00 and includes enough booklets and answer sheets to test 25 children, the material costs are not $25.00 but $25/25, or $1.00 per student. Costs associated with administration of the test are dependent upon: (1) amount of time required to give and score the test and (2) level of personnel required to give and score the test. A test that takes 1 1/2 hours to give and score and that must be given by a school psychologist whose hourly wage is $8.00 will have an administration cost of 1.5 X $8.00, or $12.00. The total costs of the test is a matter of summing both the costs involved with the testing materials and the test administration, or:

Total Cost = Cost of Materials + Cost of Administration.

Special attention should be given to both the material costs and administration costs. Often school districts will purchase materials that are reusable or that can be used with a variety of tests (e.g., #2 pencils, two-copy drawing forms). While the costs associated with these materials may seem to be extraneous to the cost of a single test, they are joint costs and should be included

in the test's material costs. One way to approach this might be to divide the costs of these materials with the total projected number of students that could realistically be expected to use the materials. It might, for example, be reasonable to anticipate that 15 children could use the same pencil but unreasonable to expect 150 children to use the same pencil. It would also be a waste of materials to discard the pencil following one evaluation. The same problem exists with instruments that require the use of fairly expensive, nonconsumable stimulus materials.

Another cost consideration is that of "surplus tests." Surplus tests include instruments that a school purchases one year knowing that all the tests will not be used during the same year to avoid the increased costs of purchasing the tests the following year. This generally would occur when the funds allocated for evaluation are not depleted and the individual involved in evaluation services wishes to use all available program funds. A different type of surplus occurs when a school obtains a special type of evaluation device that will only be used with one or two students. Although there may be enough materials to service 25 children, only a few students will be tested with the instrument, resulting in the test's maximum potential being under-utilized.

Costs associated with administration requirements also need to be carefully considered. Often educational and psychological instruments that are diagnostic in nature are untimed, allowing an examinee to work at his or her own pace. For an examinee with a specific handicap the administration time would be greater than

that expected for a non-handicapped examinee. Determining the actual time required to give the test might be best defined by the time the "average" examinee takes. Scoring time also varies from test to test. Instruments having scoring keys and templates will generally require less time to score than those not using scoring guides. Projective tests and tests that are highly subjective will require more time to score than simple objective measures. Again, the scoring time may vary from scorer to scorer; the time required by the "average" scorer might be the best estimate of scoring time for a single test. Finally, the level of personnel required to give the test and score it would influence the administration costs. Many tests are easily given and scored by a clerk while others require specially trained examiners. The former type of test would have less cost associated with personnel salaries than would the latter type.

The time required to interpret and report the results of the test is not considered here as an administration cost. Although this would certainly influence the true costs of an assessment instrument, interpretation and reporting times vary greatly from individual to individual and are very difficult to estimate.

As a next step, the combined costs of tests must be considered. Generally, all that is necessary is to compute the total cost of each test and to sum across the combination of tests. However, joint costs might be involved with some of the tests. These overlapping costs must be considered and allocated appropriately to all tests involved before an accurate assessment of the combined costs of the tests can be made (Levin, Note 5).

Once the costs associated with a single test are determined, the combined costs of the tests are defined by:

$$\text{Combined Costs} = TC_1 + TC_2 + \cdots + TC_k ,$$

where TC = total cost associated with a single test.

Cost-effectiveness Ratio. After obtaining the costs and efficiency values for a particular combination of tests, calculation of the cost-effectiveness ratio is fairly straightforward (Levin, Note 5). C/E ratios simply use the combined costs as the numerator and the total efficiency as the denominator, so that:

$$C/E = \frac{\text{Combined Costs}}{P(D/C)_1} .$$

Determining the Optimal Combination of Tests. Most handicaps or learning problems may be measured by a number of standardized instruments. It is generally not possible, however, to use all the available tests during the course of a single evaluation. The preceding cost analysis model is an attempt to devise a series of decision alternatives so that a logical choice between available tests can be made. Computing a quantity of C/E ratios for various combinations of tests is meaningless unless there is a methodological approach by which to compare them. The following two methods are presented as ways to order and interpret the obtained C/E ratios.

Descriptive approach: The first approach uses a simple descriptive graph, where test combinations are placed along the abscissa of the graph and their respective C/E ratios plotted along

the ordinate (see Figure 3). This indicates which test combination provides the maximum effectiveness per level of cost. In Figure 3, the second combination of tests appears to be the most cost-effective; that is, the costs associated with increasing a certain amount of correct classification decisions (e.g., a percentage point) are the lowest for this combination compared to the other combination of tests. This combination would be the optimal combination to choose.

Place Figure 3 About Here

Step-wise approach: A more involved approach to interpreting the C/E ratios is to determine the point at which adding more tests does not increase the C/E ratio. The combination of tests would also be placed along the abscissa of the graph and their C/E ratios plotted along the ordinate. However, the order of the combinations would progress in a logical, step-wise manner not required by the first approach. Placement on the abscissa would be based on the number of tests found in each combination of tests so that combinations containing fewer tests would be placed first with combinations containing more tests following. In cases where the same number of tests are found in more than one combination, the following order might be used:

(1) Test 1 + Test 2

(2) Test 1 + Test 3

(3) Test 1 + Test 4

(4) Test 2 + Test 3

(5) Test 2 + Test 4

(6) Test 3 + Test 4

(7) Test 1 + Test 2 + Test 3

(8) Test 1 + Test 2 + Test 4

(9) Test 1 + Test 3 + Test 4

(10) Test 2 + Test 3 + Test 4

(11) Test 1 + Test 2 + Test 3 + Test 4

An illustration of this is presented in Figure 4. The resulting function (curve) on the graph would indicate the point at which increasing the number of tests in the combination becomes meaningless; that is, the point at which adding more tests does not increase the C/E ratio. The combination of tests at that point would be the optimal combination of tests and would be selected for use.

Place Figure 4 About Here

As is obvious, with larger numbers of tests this procedure would become quite cumbersome. The introduction of mathematical procedures to order all possible combinations of tests would be of great value at this point.

Summary. The procedures presented in this paper are intended to provide a feasible alternative for those individuals involved in evaluation. Rather than rejecting or selecting tests on a subjective basis, the model offers a logical approach to the decision processes involved in evaluation.

Many variables were not considered in this model but are

important issues to address. The definition of costs needs to be explored and elaborated. Also, expansion of the model to include other evaluation activities, such as teacher reports of behavior or observations of classroom behavior, would be important. Finally, more concrete descriptions as to acceptable limits of efficiency, costs, and C/E ratios of tests and groups of tests would make the model more viable.

Reference Notes

1. Roid, G.H. Cost Analysis as a Process. Paper presented at the Conference on Cost Analysis in Evaluation, Portland, Or., February 1980.

2. Wright, A.R., Padilla, C., Renneker, E.G., & Stearns, M.S. Local Implementation of PL 94-142: Second Year Report of a Longitudinal Study. Menlo Park, Ca.: SRI International, October 1980.

3. Larsson, B. Efficiency of Some Bayesian Decision Procedures. Malmö, Sweden: Department of Educational and Psychological Research, School of Education, January 1970.

4. Brumet, M.E. Bayesian Inferential Processes in Educational Evaluation. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976.

5. Levin, H.M. Case Book on Cost Analysis in Educational Evaluation. Portland, Or.: Northwest Regional Educational Laboratory, November 1979.

References

Bible, T.D. Decision-making in public education: A production process approach for evaluating and planning school resource allocation. Unpublished doctoral dissertation, University of California, Davis, 1976.

Cronbach, L.J., & Gleser, G.C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.

Fisher, G.H. Cost considerations in systems analysis. New York: American Elsevier Publishing Company, Inc., 1971.

Jones, W.P., & Newman, F.L. Bayesian techniques for test selection. Educational and Psychological Measurement, 1971, 31, 851-856.

Kirk, S.A. Educating exceptional children (2nd ed.). Boston: Houghton Mifflin Company, 1972.

Leroy, A., & Mellergard, M. The use of information in diagnostic decisions. ACTA Psychiat. Scand., 1973, 49, 405-417.

Levin, H.M. Cost analysis. In N.L. Smith (Ed.), New techniques for evaluation. Beverly Hills, Ca.: Sage Publications, Inc., 1981.

Messick, S. Test validity and the ethics of measurement. American Psychologist, 1980, 35, 1012-1027.

Roid, G.H. Cost-effectiveness analysis in mental health policy research. American Psychologist, in press.

Schmitt, S.A. Measuring uncertainty: An elementary introduction to Bayesian statistics. Reading, Ma.: Addison-Wesley, 1969.

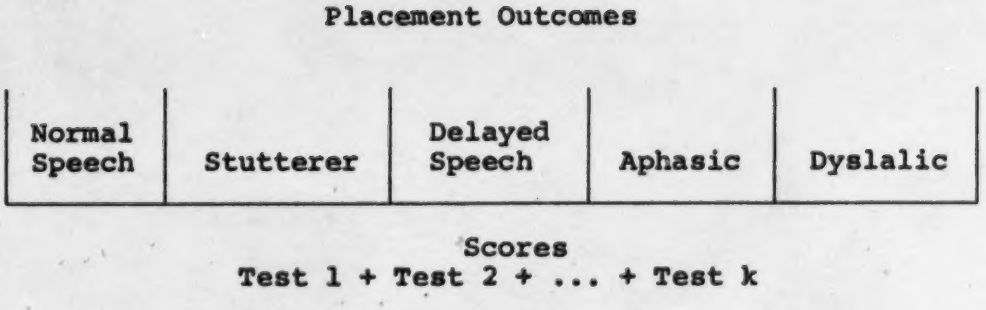

Figure 1. Illustration of a Placement Decision.

Diagnosis (based on tests)	Actual Group Membership: Subaverage (Group A)	Actual Group Membership: Average (Group B)
Subaverage (Classification A)	#1	#2
Average (Classification B)	#3	#4

Figure 2. Frequency Table for Determining P(D/C).

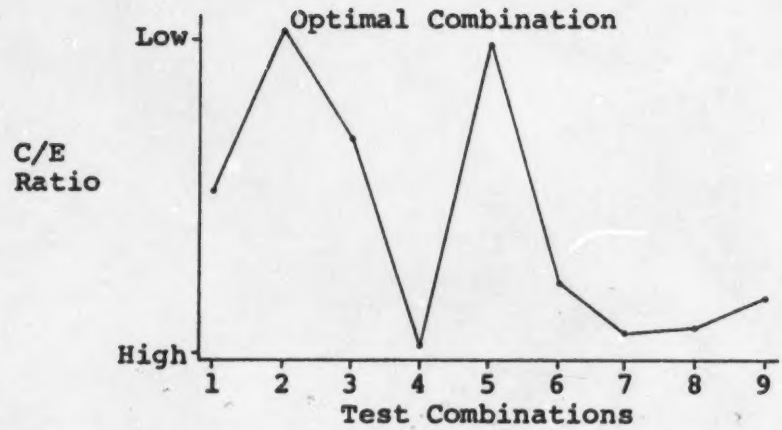

Figure 3. Descriptive Graph of C/E Ratios.

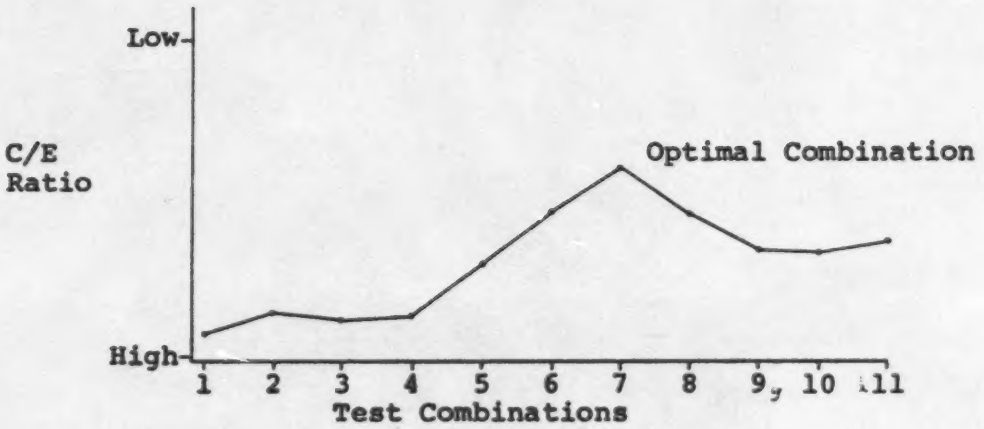

Figure 4. Step-wise Graph of C/E Ratios.

SUMMARY

$$P(C_1/D) = \frac{P(D/C_1)\ P(C_1)}{P(D)}$$

where $P(D) = P(D/C_1)\ P(C_1) + P(D/C_2)\ P(C_2)$

$P(C_1)$ = probability of belonging to Group 1

$P(C_2)$ = probability of not belonging to Group 1 but belonging to Group 2

$P(D/C_1)$ = probability of being diagnosed as belonging to Group 1, given that you actually belong to Group 1

$P(D/C_2)$ = probability of being diagnosed as belonging to Group 1, given that you actually belong to Group 2

$P(C_1/D)$ = probability of belonging to Group 1 given a diagnosis